WOODHOUSE EXHIBIT 5

EXHIBIT H

Message

From: David Esiobu [redacted]@meta.com]
Sent: 4/11/2024 10:40:44 PM
To: Nikolay Bashlykov [redacted]@meta.com]; David Esiobu [redacted]@meta.com]; Jacob Xu [redacted]@meta.com]; Frank Zhang [redacted]@meta.com]; Xiaolan Wang [redacted]@meta.com]; Viktor Kerkez [redacted]@meta.com]
Subject: Message summary [{"otherUserFbId":null,"threadFbId":7614278578629298}]

Xiaolan Wang (4/11/2024 15:40:44 PDT):
>Hi Everyone, would like to share the downloading progress of AA.
>- Internet Archive (46 TB out of 126 TB), we may get another 30TB in the next few days. However, the remaining ones will be hard to get due to too few seeds and slow downloading speed.
>- Z-library (25.7 TB out of 54 TB). We may get a few more TB in the next few days, but the remaining would be hard to acquire due to similar reasons above.
>- Libgen (10 TB out of 10 TB): we got almost all we want (all torrents posted after 2023-03-01) with a few ones pending.
>
>@Nikolay Bashlykov , I have few questions about previous libgen processing:
>1. Text ID in hive tables (llama3_scimag/llama3_scitech), is this column the md5 id you generated for deduplication? Is this based on each article/book's title extracted from source files?
>2. For data processing (https://github.com/fairinternal [redacted] I noticed that you performed content removal with a threshold, I would like to get some intuition on this. Is goal to remove things like table of contents at the beginning and references at the end?